Vlerick Leuven Gent Management School

the Autonomous Management School of
Ghent University and Katholieke Universiteit Leuven

MASTERS IN BUSINESS ADMINISTRATION
2011 - 2012

INDIVIDUAL CRITICAL PAPER:
THE SUCCESS OF DIFFERENT DISCOUNTS ON OXYPASS

Fundamentals of Business Statistics

Prof Dr Ir. Hendrik Vanmaele
Prof. Dr. Bram Desmet

Pascal Vyncke

Engaging in irregularities is severely sanctioned in correspondence with article 34 of the Examination rules.

We hereby declare that we have not engaged in any such irregularities.

Student(s)'s signature(s)

Table of contents

Introduction 2
The dataset 4
Variables description 4
Statistics 4
Histogram 5
Box plot 5
Correlation 6
Scatterplot 7
Confidence interval 7
Hypothesis test 8
Multiple Linear Regression 8
Annex: the dataset 10

Introduction

This individual critical paper for Business Statistics is analysing a dataset from our company. SeniorenNet has a product called OxyPass. This is a pass that not only supports the free website SeniorenNet, but also gives discounts to his owners. The usage of the online discounts is analysed in this paper.
Several techniques are used like descriptive statistics, histogram, box plot, correlation, hypothesis test and linear regression.

Bold and colours like red and orange are used to increase the readability of the paper.

The dataset

The dataset (in annex) consists of 315 data points and 5 columns. The data is the usage of all the **online discounts by the customers** of OxyPass.be, with discounts at least 3 months in usage. People buy an OxyPass and can get discounts with it.

Variables description

The SectionID and ItemID are **nominal variables**. The Average Usage Per Week is a **continuous** variable. The Number of Uses and Days Online are both **discrete** variables.

Statistics

Below are descriptive statistics for the three most important variables.
The ones in **red** are most relevant and **I will comment on these**.

One Variable Summary	Number of Uses	Days Online	Average Usage Per Week
Mean	51,70	227,34	1,452
Variance	8821,18	4124,77	7,279
Std. Dev.	93,92	64,22	2,698
Skewness	3,0190	-0,3747	3,3426
Kurtosis	13,6593	1,5450	16,0387
Median	10,00	250,00	0,326
Mean Abs. Dev.	60,89	59,70	1,670
Mode	1,00	295,00	0,039
Minimum	1,00	100,00	0,024
Maximum	642,00	296,00	18,760
Range	641,00	196,00	18,736
Count	315	315	315
Sum	16284,00	71613,00	457,227
1st Quartile	3,00	171,00	0,122
3rd Quartile	55,00	289,00	1,405
Interquartile Range	52,00	118,00	1,283

Number of Uses

For the Number of Uses is the mean (51,70) and median (10) **very different**. This is confirmed by the **high range** (641) and the **high standard deviation** (93). This gives us the knowledge that there is a lot of difference in the use of discounts: some are used heavily, others are not used at all. All together the different discounts are used 16.284 times in the 10 months of data we have.

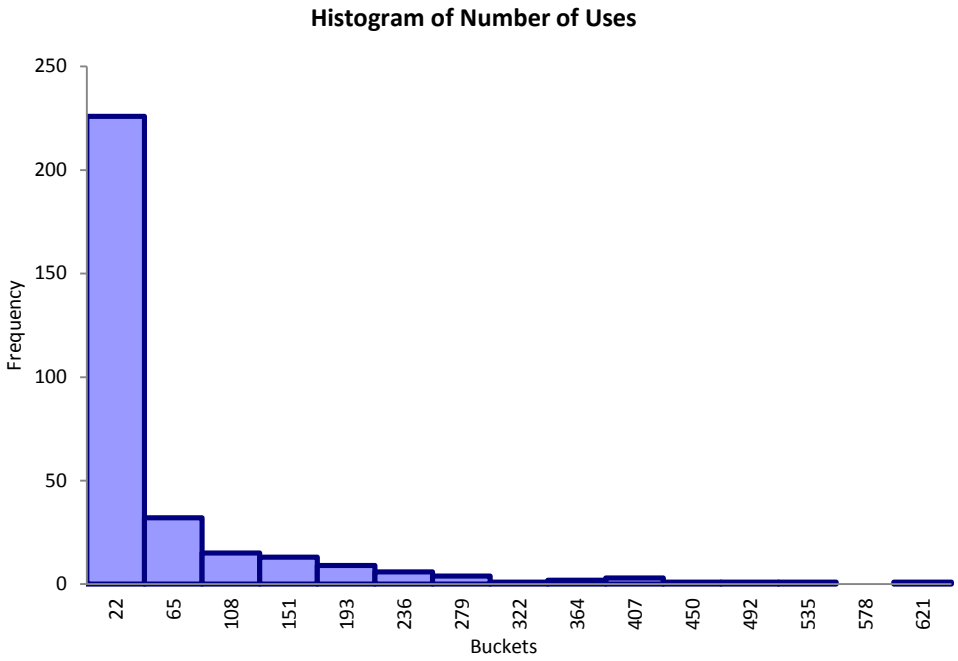
Days Online

When we look to the Days Online we see that the **mean** of days online is 227 days and this is **not very different from the median** of 250 days. Still we see that with the range of 196 days and an interquartile range of 118 days there is still a lot of difference of almost 4 months between the first and third quartile.

Average Usage Per Week

I also looked to the variable of Average Usage Per Week. This ignores the fact that some discounts are longer available than others and really looks to the average usage. Here we see that the highest (18,760) and lowest (0,024) number are very much apart from each other. We see with the standard deviation of 2,69 that there is also **a lot of difference in usage**, even if we look to the average usage per week.

Histogram

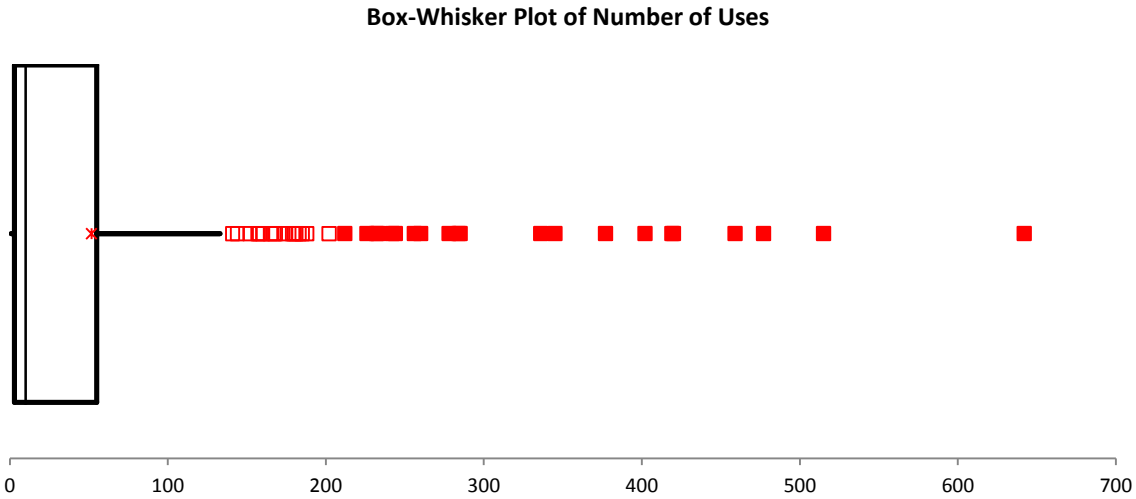


When we look to the histogram of the number of uses, it is clear that **most of the discounts were used less than 22 times**. We can see in this histogram why we have a very different mean (51) and median (10). Most numbers are less than 22 which gives us the low median. But because there are also several very high usages, this drives the mean to a higher number. This also **explains the high standard deviation** (93): a lot is much lower than the mean (the mode is even 1,00), while also several are much higher than the mean (as much as 642).

Box plot

The following box plot is made for the Number of Uses. We see a visual representation what we already could see in the descriptive statistics calculated before.

The first quartile is very small (3) and tells us that 25% of the discounts is almost not used. The mean (51) is almost the same as the 3rd quartile (55). We can see that although the mean is 51, almost 75% of the usage is lower than the mean (to be correct: lower than 55). This suggests that there are **very high outliers that increase the mean**. We can see this on the box-plot diagram with several extreme outliers to the right.



We learn from this that **a lot** of discounts are almost **not used**, but that there are **several** discounts that are **heavily** used; while there is a big chunk that is used once in a while.

This is **consistent with the original business model**: get a few extreme good discounts which can't be found anywhere else. To be complete we also put 'general' discounts together (and these general discounts are not that extreme, and it looks that they are used less).

Correlation

The following table gives the correlation between all used variables.

Correlation Table	Item ID	Number of Uses	Section ID	Days Online	Average Usage Per Week
Item ID	1,000	-0,460	-0,735	-0,547	-0,423
Number of Uses	-0,460	1,000	0,403	0,302	0,959
Section ID	-0,735	0,403	1,000	0,277	0,381
Days Online	-0,547	0,302	0,277	1,000	0,185
Average Usage Per Week	-0,423	0,959	0,381	0,185	1,000

Expected correlations

There are some higher correlations which can be explained:

- The fact that SectionID and ItemID have a high negative correlation (-0.7) is because the fact that with the start-up of the project items where **added section by section**. The correlation is not 1 because after the start-up items are added on a need basis and this can be any section.

- A higher negative correlation of Item ID and Days Online (-0.54) is also explained by the fact that ID's are just **serial numbers**, lower ID's are used earlier (so higher Days Online). The correlation is also not maximum because not all ID's are used and some are not in use anymore (or items are taken offline for a short time).

- The very high correlation between the Average Usage Per Week and Number of Uses (0,95) is because of the fact that **our system** probably **uses** the Number of Uses to calculate the Average Usage Per Week.

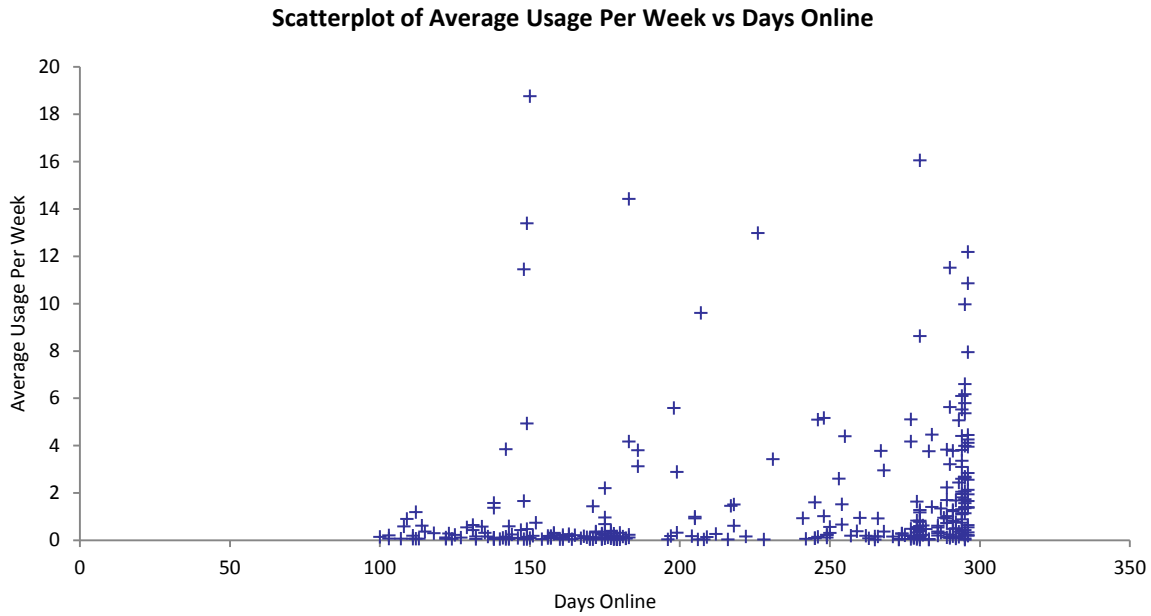
This is new!

Surprising for me was that the correlation of Number of Uses and Days Online is somewhat low (0,381). And even more surprising is the fact that the **Average Usage Per Week has almost no correlation with Days Online** (0,185).

Our company generally assumed that the longer a discount was online, the more it was used because it gets more widely known.

Scatterplot

Because of this new insight I draw a scatterplot of the Average Usage Per Week versus the Days Online.



Correlation confirmed

I can see indeed that there is **almost no connection** between the Average Usage Per Week and the Days Online. You can see extreme outliers on almost all different days online. This explains the very low correlation.

But...[1]

There is one thing that I can see on the scatterplot and that is that there are no outliers the first 150 days. I can also see that the **oldest discounts** do have some **more extreme outliers** than the younger discounts. This suggests that there is some kind of correlation to the age of discounts: the very young are indeed not so popular, the more older ones have higher peaks of high usage.

Confidence interval

An estimation for a 95% confidence interval for the mean of the Average Usage Per Week is [1.152, 1.751]. This means we have a 95% confidence that if we take a sample of our discounts, the mean will be between 1.152 and 1.751 uses per week (**roughly 60 to 90 uses per year**). This is interesting to know to communicate to future advertisers.

By comparing this interval to the median (0.32), we can see that the extreme outliers have an impact to increase the mean.

[1] Text in orange is interpretation that is potentially risky and can be false.

Hypothesis test

We make a hypothesis test where we look to the (alternative) hypothesis that an advertiser that gives a discount, gets **at least 1 customer per week** (4 customers per month) extra. The null hypothesis is that he gets less than 1 customer per week extra. This is a one-tailed test.

We get the following calculations:

Hypothesis Test (One-Sample)	Average Usage Per Week
Sample Size	315
Sample Mean	1,452
Sample Std Dev	2,698
Hypothesized Mean	1
Alternative Hypothesis	> 1
Standard Error of Mean	0,152
Degrees of Freedom	314
t-Test Statistic	2,9703
p-Value	0,0016
Null Hypoth. at 10% Significance	Reject
Null Hypoth. at 5% Significance	Reject
Null Hypoth. at 1% Significance	Reject

According to this results we should reject with a very high certainty the null hypothesis. We can conclude that the chance that you get less than 1 customer per week as a mean is very unlikely (with a p-value of 0,0016). The very low p-value gives us a very high confidence and **strong evidence** for this alternative hypothesis.

Multiple Linear Regression

I want to get to a formula that predicts the Number of Uses (dependent value), using the other variables.
When using only the independent value DaysOnline I get the following results.

Summary	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate
	0,3022	0,0913	0,0884	89,67189657

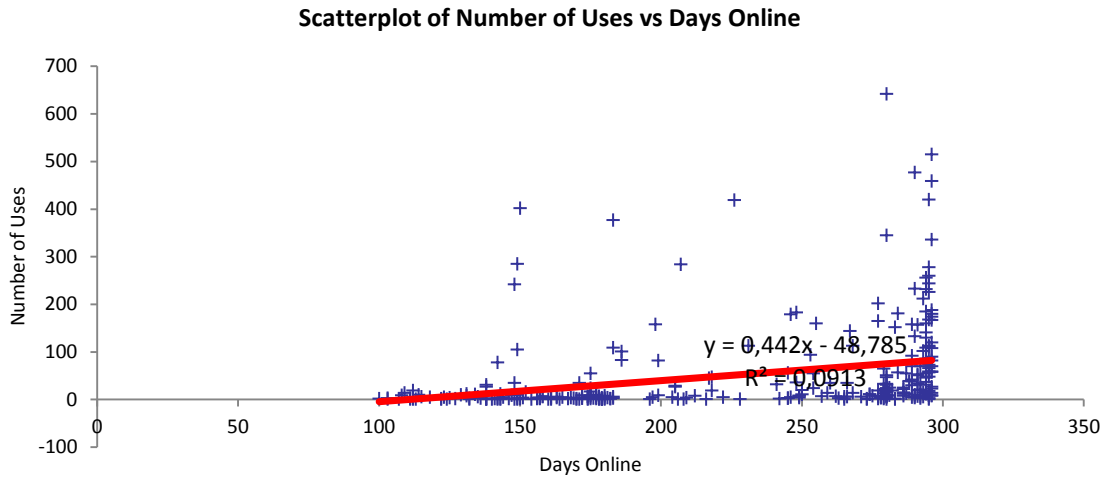
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio	p-Value
Explained	1	253002,3952	253002,3952	31,4639	< 0.0001
Unexplained	313	2516848,348	8041,049034		

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Confidence Interval 95% Upper
	-				-	-
Constant	48,78463389	18,61208696	-2,6211	0,0092	85,40525542	12,16401235
Days Online	0,441975056	0,078793755	5,6093	< 0.0001	0,286942668	0,597007444

This gives the following formula:

NumberOfUses = 48,78 + 0,44 DaysOnline

I now also add the regression line to the scatterplot (in red).



This is only with a R-Square of 0,09 and is very low and a high standard deviation of the (regression) error (89,6). Therefore I go to **multiple variables**.

Using the Days Online and Average Usage Per Week as independent values I get the following results.

Summary	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate
	0,9676	0,9363	0,9359	23,78756308

ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio	p-Value
Explained	2	2593306,118	1296653,059	2291,5212	< 0.0001
Unexplained	312	176544,6251	565,8481575		

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
	-				-	-
Constant	38,66062342	4,939799404	-7,8264	< 0.0001	48,38015542	28,94109141
Days Online	0,189562464	0,021267187	8,9134	< 0.0001	0,147717221	0,231407707
Average Usage Per Week	32,55924233	0,506276922	64,3111	< 0.0001	31,56309363	33,55539103

This gives the following formula:

NumberOfUses = 38,67 + 0,19 DaysOnline + 32,56 AverageUsagePerWeek

With an **explanation of 93%** of the variation of NumberOfUses this formula gives a **good prediction of the value** (although still a standard deviation of the error of 23,78, but almost 4 times lower than before). Using the values DaysOnline and AverageUsagePerWeek also looks very useful: the p-values are extremely small.

So it is possible to predict the number of uses of a discount if we know how long a discount is online and what the average usage per week is. This can be used to test a new discount during a short period (then we know the average usage per week). After this period we can give a number of uses on any date we want (knowing that R Square is 93%).

Vlerick Leuven Gent Management School

the Autonomous Management School of
Ghent University and Katholieke Universiteit Leuven

Annex: the dataset

This is the data used for the analysis. **Please note, data is confidential.**

Item ID	Number of Uses	Section ID	Days Online	Average Usage Per Week
4	6	6	290	0,14
5	23	6	296	0,54
6	65	2	279	1,63
7	7	2	278	0,18
8	244	3	295	5,79
9	113	6	295	2,68
10	160	3	294	3,81
12	31	3	294	0,74
13	3	6	292	0,07
14	185	5	294	4,40
15	86	6	294	2,05
16	167	5	296	3,95
17	256	3	294	6,10
18	48	3	295	1,14
19	515	5	296	12,18
20	59	5	295	1,40
21	120	5	296	2,84
22	19	5	294	0,45
23	459	3	296	10,85
24	39	5	288	0,95
25	70	3	296	1,66
26	55	6	287	1,34
27	87	5	295	2,06
28	37	5	294	0,88
29	336	3	296	7,95
30	76	2	294	1,81
31	8	3	275	0,20
32	113	6	268	2,95
33	212	3	293	5,06
34	24	6	254	0,66
35	20	2	250	0,56
36	21	3	292	0,50
37	102	3	293	2,44
38	158	3	289	3,83
39	188	3	296	4,45
40	22	3	279	0,55
41	55	3	294	1,31
42	174	3	296	4,11
43	33	2	280	0,83
44	180	3	296	4,26
45	233	3	290	5,62
46	92	3	289	2,23
47	109	3	294	2,60
48	23	3	280	0,58
49	38	3	289	0,92
50	420	3	295	9,97
51	82	3	294	1,95
52	260	3	295	6,17
53	58	2	296	1,37
54	168	2	295	3,99
55	31	3	280	0,78
56	70	3	292	1,68
57	133	3	290	3,21
58	18	3	295	0,43
59	278	3	295	6,60
60	157	3	291	3,78
61	82	3	296	1,94
62	51	3	291	1,23
63	53	3	291	1,27
64	477	3	290	11,51
65	57	3	284	1,40
66	19	3	282	0,47
67	642	3	280	16,05
68	165	2	277	4,17
69	160	3	255	4,39
70	25	2	286	0,61
71	179	3	246	5,09
72	183	3	248	5,17
73	8	2	280	0,20
74	202	2	277	5,10
75	113	3	231	3,42
76	36	2	248	1,02
77	419	3	226	12,98
78	9	3	199	0,32
79	19	3	218	0,61
80	284	3	207	9,60
81	101	3	186	3,80
82	27	3	205	0,92
83	82	3	199	2,88
84	29	3	205	0,99
85	5	3	197	0,18
86	158	3	198	5,59
87	83	3	186	3,12
89	55	3	175	2,20
90	35	3	171	1,43
91	402	3	150	18,76
92	105	3	149	4,93
93	285	3	149	13,39
94	35	3	148	1,66
95	242	3	148	11,45

96	12	3	143	0,59	365	15	1	280	0,38
97	78	3	142	3,85	371	33	1	279	0,83
98	12	3	131	0,64	372	2	1	283	0,05
99	27	3	138	1,37	375	5	1	123	0,28
100	31	3	138	1,57	382	1	1	283	0,02
101	11	3	134	0,57	384	7	1	280	0,18
102	14	3	109	0,90	385	1	1	295	0,02
103	9	3	108	0,58	388	12	1	294	0,29
104	10	3	114	0,61	562	42	1	289	1,02
105	19	3	112	1,19	563	11	1	250	0,31
157	5	1	204	0,17	564	12	1	286	0,29
158	152	1	283	3,76	565	10	1	295	0,24
159	66	1	295	1,57	566	8	1	249	0,22
160	130	1	294	3,10	567	4	1	278	0,10
164	13	1	279	0,33	568	4	1	249	0,11
168	14	1	296	0,33	569	19	1	277	0,48
169	21	1	279	0,53	570	1	1	112	0,06
172	141	1	294	3,36	571	6	1	278	0,15
175	70	1	289	1,70	572	19	1	279	0,48
181	51	1	280	1,28	573	7	1	277	0,18
182	6	1	293	0,14	575	181	1	284	4,46
183	9	1	295	0,21	576	1	1	280	0,03
186	48	1	294	1,14	577	3	1	263	0,08
187	73	1	295	1,73	578	2	1	277	0,05
189	1	1	216	0,03	579	5	1	289	0,12
207	111	1	295	2,63	580	3	1	245	0,09
210	47	1	280	1,18	581	20	1	291	0,48
211	8	1	249	0,22	582	4	1	278	0,10
213	9	1	280	0,23	583	30	1	289	0,73
215	11	1	274	0,28	584	4	1	279	0,10
216	34	1	291	0,82	585	6	1	292	0,14
220	3	1	295	0,07	586	4	1	156	0,18
223	10	1	290	0,24	587	9	1	295	0,21
224	144	1	267	3,78	588	6	1	265	0,16
225	27	1	296	0,64	589	6	1	271	0,15
231	22	1	286	0,54	590	7	1	262	0,19
244	108	1	296	2,55	591	1	1	107	0,07
245	226	1	295	5,36	593	9	1	281	0,22
258	14	1	259	0,38	595	9	1	296	0,21
267	7	1	291	0,17	596	11	1	274	0,28
271	5	1	280	0,13	597	3	1	245	0,09
287	7	1	257	0,19	599	24	1	280	0,60
297	17	1	294	0,40	600	2	1	279	0,05
312	1	1	208	0,03	601	1	1	273	0,03
314	4	1	209	0,13	602	18	1	280	0,45
319	1	1	196	0,04	603	2	1	273	0,05
322	345	1	280	8,63	605	32	1	292	0,77
324	8	1	296	0,19	606	4	1	290	0,10
329	15	1	293	0,36	607	5	1	222	0,16
331	59	1	296	1,40	608	8	1	286	0,20
333	1	1	178	0,04	612	13	1	287	0,32
336	7	1	280	0,18	613	5	1	246	0,14
338	57	1	295	1,35	616	232	1	294	5,52
342	16	1	280	0,40	617	46	1	294	1,10
348	25	1	281	0,62	618	90	1	296	2,13
349	15	1	286	0,37	619	4	1	163	0,17
356	1	1	265	0,03	620	5	1	175	0,20
360	1	1	206	0,03	621	17	1	175	0,68

622	9	1	147	0,43
623	9	1	175	0,36
624	1	1	160	0,04
625	24	1	175	0,96
626	109	1	183	4,17
627	14	1	268	0,37
628	1	1	124	0,06
629	8	1	212	0,26
632	6	1	158	0,27
638	2	1	179	0,08
639	4	1	181	0,15
640	6	1	266	0,16
642	6	1	183	0,23
643	5	1	118	0,30
644	3	1	183	0,11
645	10	1	177	0,40
646	3	1	180	0,12
647	35	1	260	0,94
648	6	1	135	0,31
649	1	1	179	0,04
651	1	1	161	0,04
652	6	1	174	0,24
653	1	1	154	0,05
654	8	1	172	0,33
655	5	1	165	0,21
656	3	1	172	0,12
657	35	1	266	0,92
660	3	1	149	0,14
661	1	1	174	0,04
662	2	1	138	0,10
664	7	1	178	0,28
667	1	1	164	0,04
669	5	1	178	0,20
671	2	1	242	0,06
672	45	1	217	1,45
673	3	1	136	0,15
674	4	1	125	0,22
676	10	1	149	0,47
678	2	1	127	0,11
679	9	1	172	0,37
680	3	1	172	0,12
681	2	1	182	0,08
685	1	1	138	0,05
688	2	1	100	0,14
691	1	1	140	0,05
694	3	1	111	0,19
696	2	1	146	0,10
697	6	1	163	0,26
701	2	1	156	0,09
705	5	1	161	0,22
706	8	1	131	0,43
707	9	1	177	0,36
708	3	1	157	0,13
711	1	1	113	0,06
712	55	1	254	1,52
713	94	1	253	2,60
714	2	1	170	0,08
719	1	1	122	0,06
722	56	1	245	1,60
723	32	1	241	0,93
724	1	1	138	0,05
726	47	1	218	1,51
727	1	1	228	0,03
728	1	1	148	0,05
729	1	1	170	0,04
730	4	1	178	0,16
732	4	1	168	0,17
734	1	1	149	0,05
735	2	1	175	0,08
736	2	1	157	0,09
737	2	1	179	0,08
739	2	1	178	0,08
740	2	1	138	0,10
741	3	1	178	0,12
742	4	1	177	0,16
744	2	1	177	0,08
745	4	1	179	0,16
751	2	1	141	0,10
752	3	1	161	0,13
755	2	1	167	0,08
757	1	1	111	0,06
759	1	1	179	0,04
762	1	1	180	0,04
764	1	1	171	0,04
766	1	1	179	0,04
769	2	1	143	0,10
770	1	1	143	0,05
773	2	1	150	0,09
775	1	1	103	0,07
785	2	1	174	0,08
787	1	1	142	0,05
798	2	1	182	0,08
799	4	1	151	0,19
815	7	1	174	0,28
816	3	1	142	0,15
822	1	1	132	0,05
823	377	1	183	14,42
824	8	1	180	0,31
825	3	1	176	0,12
826	5	1	144	0,24
827	3	1	132	0,16
828	7	1	158	0,31
829	3	1	169	0,12
832	4	1	157	0,18
833	16	1	152	0,74
835	10	1	129	0,54
845	3	1	103	0,20
846	2	1	122	0,11
852	6	1	115	0,37